PR #42126 完整报告

vllm-project/vllm

[CI][AMD] Skip tests where models have problems or fails on both HW types

合并时间：2026-05-14 16:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42126

执行摘要

- 一句话：跳过 ROCm 上已知有问题的多模态测试
- 推荐动作：值得合入以保持 CI 稳定性。PR 本身是临时缓解措施，建议跟踪关联 issue 并在上游模型修复后移除这些跳过标记。

功能与动机

ROCm 上运行多模态生成测试时，rhymes-ai/Aria 需要更新 transformers 以兼容 vision_processor.py（issue）；zai-org/glm-4v-9b 也存在兼容性问题（discussion）；zai-org/GLM-OCR 在 H100 和 MI300 上均输出乱码（issue）。为保障 CI 稳定性，临时跳过这些失败的测试。

实现拆解

1. 定位测试配置：在文件 tests/models/multimodal/generation/test_common.py 中找到 VLMTestInfo 字典，该字典定义了每个多模态模型的测试参数。
2. 为三个模型添加 pytest.mark.skip：在 Aria、GLM-4V 和 GLM-OCR 的 marks 列表中各添加了一个 pytest.mark.skip 标记，附带明确的跳过原因和关联 issue/ 讨论链接。
3. 保留原有的 GPU 内存标记：原有的 large_gpu_mark 标记保留在 marks 列表中，确保跳过条件仍包含硬件资源约束。

关键文件：

- tests/models/multimodal/generation/test_common.py（模块 测试配置；类别 test；类型 test-coverage）：多模态生成测试的配置文件，为三个已知有问题的模型添加了 pytest.mark.skip 标记，避免 CI 被不相关的失败干扰。

关键符号：未识别

关键源码片段

tests/models/multimodal/generation/test_common.py

多模态生成测试的配置文件，为三个已知有问题的模型添加了 pytest.mark.skip 标记，避免 CI 被不相关的失败干扰。

```
# tests/models/multimodal/generation/test_common.py
# aria 模型：需要更新 transformers 以支持 vision_processor.py
```

```
"aria": VLMTestInfo(
    models=["rhymes-ai/Aria"],
    test_type=(VLMTestType.IMAGE, VLMTestType.MULTI_IMAGE),
    # ... 其他配置 ...
    marks=[
        pytest.mark.skip(
            reason="Aria needs to update for latest transformers, "
                   "must have a vision_processor.py. "
                   "An issue has been filed: "
                   "https://huggingface.co/rhymes-ai/Aria/discussions/23"
        ),
        large_gpu_mark(min_gb=64),
    ],
),
# glm4v 模型：代码存在 bug
"glm4v": VLMTestInfo(
    models=["zai-org/glm-4v-9b"],
    test_type=VLMTestType.IMAGE,
    # ... 其他配置 ...
    marks=[
        pytest.mark.skip(
            reason="The code for this model has a bug. "
                   "Please see the issue here: "
                   "https://huggingface.co/zai-org/glm-4v-9b/discussions/46"
        ),
        large_gpu_mark(min_gb=32),
    ],
),
# GLM-OCR 模型：在 AMD 和 NV 上均失败
"glm_ocr": VLMTestInfo(
    models=["zai-org/GLM-OCR"],
    # ... 其他配置 ...
    marks=[
        pytest.mark.skip(
            reason="This test fails on both AMD and NV "
                   "hardware. Please see the issue: "
                   "https://github.com/vllm-project/vllm/issues/42016"
        ),
        large_gpu_mark(min_gb=32),
    ],
),
```

评论区精华

gemini-code-assist[bot] 自动审查发现三个 pytest.mark.skip 的 reason 字符串存在拼接空格缺失问题，导致链接与文字粘连（如 'filed:https://...'、'bug.Please'），并建议使用 reason= 关键字参数以保持一致性。作者 rasmith 回复 "Done" 并修复了这些问题。最终批准者 tjtanaa 表示 LGTM。

- skip 标记字符串格式问题 (style): 作者 rasmith 接受建议并修复了格式问题。

风险与影响

- 风险：无技术风险。变更仅限于测试跳过标记，不涉及任何生产代码、模型权重或推理逻辑。跳过测试的模型本身存在问题，跳过不会掩盖有效的失败。
- 影响：仅影响 CI 测试流程。三个已知有问题的模型测试将被跳过，减少 ROCm CI 的噪音和失败次数，但对用户功能无任何影响。其他多模态测试继续正常运行。
- 风险标记：测试覆盖调整

关联脉络

- 暂无明显关联 PR